Using Web-scale N-grams to Improve Base NP Parsing Performance

Emily Pitler
University of Pennsylvania

Shane Bergsma
University of Alberta

Dekang Lin
Google, Inc.

Kenneth Church
Johns Hopkins University

COLING 2010
Overview

- **Goal:** Recover the syntactic structure of arbitrary noun phrases (NPs):
  - its remaining mortgage loan origination operations
  - \((\text{its} \ (\text{remaining} \ (((\text{mortgage} \ \text{loan}) \ \text{origination}) \ \text{operations}))))\)

- **Method:** score all spans with classifier
- **Features:** statistics from web-scale N-grams
- **Result:** 95.4% exact match accuracy
Parsing NPs

• Classic 3-word cases: Left or Right?

- (social science) teacher
- retired (science teacher)

• Our objective: NPs of arbitrary length, including conjunctions:
  (French ((television and movie) producers))
Parsing NPs

• It’s Hard: Number of binary trees increases with the Catalan numbers [Church & Patil, 1982]:
  2 (three words), 5, 14, 42 (six words), 132, 429, …

• It’s Worth Doing:
  70% of search-engine queries [Barr et al. 2008]
  (washed baby) carrots vs. washed (baby carrots)
Annotated Data

- Base NPs: NPs with no embedded NPs
- Penn Treebank originally had flat base NPs:
  (NP (JJ time-limited) (NN poison) (NNS pills))
- [Vadas and Curran, 2007] provide annotations for NPs in the WSJ articles from the Penn Treebank
  – 98.5% inter-annotator agreement accuracy
Examples

(( (biotechnology research) (and (vaccine manufacturing)) ) concerns)
((The government)'s)
(executive (vice president))
(the (first time))
(its (four-year history))
(an (adverse (net-benefit decision)))
(the (same conclusions))
(industry (,(science (and technology))))
(Merieux (and Connaught))
(early (this week))
(the (government decision))
(Mehta (,&lsaly))
((the government)'s)
(an (out-of-court settlement))
(a (settlement proposal))
(((research (and development)) spending) levels)
(Toronto-based (((Richardson Greenshields) Inc.))))
Method: Span Scoring

• Supervised classifier (SVM) predicts probability of each span:
  
  \( (\text{French television}) \) and movie producers
  
  \( \text{French (television and)} \) movie producers
  
  \( (\text{French television and}) \) movie producers
  
  ... 

  \( \text{French television and (movie producers)} \)

• Can be feature in structured predictor

  [Taskar et al. 2004]
Method: Span Scoring

French television and movie producers

<table>
<thead>
<tr>
<th>(French television)</th>
<th>French</th>
<th>television</th>
<th>and</th>
<th>movie</th>
<th>producers</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>.15</td>
<td>.07</td>
<td>.37</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>television</td>
<td></td>
<td>.14</td>
<td>.51</td>
<td>.76</td>
<td></td>
</tr>
<tr>
<td>and</td>
<td></td>
<td></td>
<td>.78</td>
<td>.33</td>
<td></td>
</tr>
<tr>
<td>movie</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.25</td>
</tr>
<tr>
<td>producers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(French ((television (and movie)) producers)))
Prior work leveraging raw text

“retired science teacher”

- **Adjacency model** [Marcus, 1980; Liberman and Sproat, 1992; Pustejovsky et al., 1993; Resnik, 1993]
  - “retired science” vs. “science teacher” (e.g. PMI)

- **Dependency model** [Lauer, 1995]
  - “retired science” vs. “retired teacher” (e.g. PMI)

- We generalize these models for features for longer NPs (counts from web-scale N-grams)
PMIs for “movie producers”

e.g. “French television and movie producers”

• PMI used as association measure

\[
PMI(x, y) = \log \frac{p(x, y)}{p(x)p(y)}
\]

• Features for all pairs:
  a) High PMI(movie, producers) \(\rightarrow\) positive evidence
  b) ↑PMI(and, movie) \(\rightarrow\) negative evidence
  c) ↑PMI(television, and) \(\rightarrow\) slightly-positive evidence
## PMIs for “movie producers”

<table>
<thead>
<tr>
<th>French</th>
<th>television</th>
<th>producers</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>and</td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>movie</td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>producers</td>
<td></td>
</tr>
<tr>
<td>television</td>
<td>and</td>
<td></td>
</tr>
<tr>
<td>television</td>
<td>movie</td>
<td></td>
</tr>
<tr>
<td>television</td>
<td>movie</td>
<td></td>
</tr>
<tr>
<td>television</td>
<td>and</td>
<td></td>
</tr>
<tr>
<td>television</td>
<td>and</td>
<td>producers</td>
</tr>
</tbody>
</table>
Special Conjunction Features

• How to tell “television and movie” go together?
  – $\text{PMI}(\text{television, and}), \text{PMI}(\text{and, movie}), \text{PMI}(\text{television, movie})$ are insufficient

• Special $\text{PMI}_{\text{and}}$ features:
  $\text{PMI}_{\text{and}}(\text{television, movie}) = \log\left(\frac{p(\"television and movie\")}{p(\"television and\")p(\"and movie\")}\right)$
Method: Span Scoring

French television and movie producers

(French ((television (and movie)) producers)))
Web-Scale N-gram Data

- Details in: [Lin et al., LREC 2010]
  - Same source as Google N-grams Version 1
  - More pre-processing: duplicate sentence removal, length+alphabetical constraints
  - Fast lookup tools based on suffix arrays

```
time cat phrases.txt | multi_lookup /export/ws09/dlin/church/Google/V2
real    21m50.273s
user    0m1.131s
sys     0m46.021s
```

168K phrases
Non N-gram Features

• Lexical features
  – Word at each position

• “Shape” features
  – Captures upper and lower case, punctuation

• Position feature
  – Prior probability of bracketing at that position
Experiments

- Standard splits of WSJ data, using annotations from [Vadas & Curran, 2007]
- All >2-word Base NPs
- Report accuracy (%)
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right-bracketing baseline</td>
<td>72.6</td>
</tr>
<tr>
<td>Lex, Shape, Position Features</td>
<td>94.0</td>
</tr>
<tr>
<td>+N-gram PMI features</td>
<td>95.4</td>
</tr>
<tr>
<td>[Vadas &amp; Curran, 2007b]</td>
<td>93.0*</td>
</tr>
</tbody>
</table>

*not really comparable
Accuracy by NP length

Accuracy
Chance (1/#Binary Trees)
Right-bracketing

Accuracy
Chance (1/#Binary Trees)
Right-bracketing
N-gram data helps most on conjunctions

<table>
<thead>
<tr>
<th>Method</th>
<th>Conjunctions</th>
<th>Everything else</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lex, Shape, Position</td>
<td>84.0</td>
<td>94.5</td>
</tr>
<tr>
<td>+Ngrams</td>
<td>89.7 (+6%)</td>
<td>95.7 (+1%)</td>
</tr>
</tbody>
</table>
## Less Data is Worse Data

<table>
<thead>
<tr>
<th>N-gram Cut-off</th>
<th># Unique N-grams</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>4,145,972,000</td>
<td>95.40%</td>
</tr>
<tr>
<td>100</td>
<td>391,344,991</td>
<td>95.30%</td>
</tr>
<tr>
<td>1,000</td>
<td>39,368,488</td>
<td>95.20%</td>
</tr>
<tr>
<td>10,000</td>
<td>3,924,478</td>
<td>94.80%</td>
</tr>
<tr>
<td>100,000</td>
<td>386,639</td>
<td>94.80%</td>
</tr>
<tr>
<td>1,000,000</td>
<td>37,567</td>
<td>94.40%</td>
</tr>
<tr>
<td>10,000,000</td>
<td>3,317</td>
<td>94.00%</td>
</tr>
</tbody>
</table>
What about other data?

Billions of Tokens

- NEWS
- Google V2
- Google V1
Number of Unique N-grams!
Number of Unique N-grams!

Web gets bigger + thresholds get lower
Conclusion

• New standard in Base NP parsing performance: 95.4%
• N-gram data particularly helps on conjunctions
• Log-linear gain with amount of UNLABELED data (number of unique N-grams)
Thanks

• Center for Language & Speech Processing at Johns Hopkins University
• Our colleagues on Team N-gram